Check Plagiarism Detection for Indonesian Language.pdf

by Arta Sundjaja

Submission date: 10-May-2019 05:09PM (UTC+0700)
Submission ID: 1102514324
File name: Plagiarism_Detection_for_Indonesian_Language.pdf (908.62K)
Word count: 2886
Character count: 15249

Journal of Physics: Conference Series

Plagiarism Detection for Indonesian Language using Winnowing with Parallel Processing

Plagiarism Detection for Indonesian Language using Winnowing with Parallel Processing

Y Arifin[1,3], S M Isa[2], L A Wulandhari[3] and E Abdurachman[2]

[1]Computer Science Department, BINUS Graduate Program – Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
[2]Computer Science Department, BINUS Graduate Program – Master of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480
[3]Computer Science Department, School of Computer Science, Bina Nusantara University, Jakarta, Indonesia 11480

Email: yarifin@binus.edu, sisa@binus.edu, lwulandhari@binus.edu, edia@binus.edu

Abstract.The plagiarism has many forms, not only copy paste but include changing passive become active voice, or paraphrasing without appropriate acknowledgment. It happens on all language include Indonesian Language. There are many previous research that related with plagiarism detection in Indonesian Language with different method. But there are still some part that still has opportunity to improve. This research proposed the solution that can improve the plagiarism detection technique that can detect not only copy paste form but more advance than that. The proposed solution is using Winnowing with some addition process in pre-processing stage. With stemming processing in Indonesian Language and generate fingerprint in parallel processing that can saving time processing and produce the plagiarism result on the suspected document.

1. Introduction

Searching technology is growing and increasing easily to get any information on the internet. Through a search website like Google makes it easy for anyone to get good information in the form of news, science, results of assignments and others. Computers and the Internet are the main cause of increased plagiarism.[1]. Plagiarism in general has several meanings: acknowledging the work of others or the ideas of others without rewarding or citation referring to the producer of the work, giving false information about the reference source, changing some words from a series of sentences from the work of others without give rewards [2]. The plagiarism can happen in any context such at the industry or at the academy, or can be in documents or in programming or other context [3]. According to the Pew Research Center study in 2011, there is an increasing in cyber plagiarism among students, it is around 55% of students admitted to plagiarism in the work of scientific tasks. Not only among the students, but also among researchers also do plagiarism in making scientific writing [4].

M. Alzahrani divides plagiarism into two type, the first one is literally plagiarism by adding or changing similar words and the second one is intelligently plagiarism by doing paraphrasing or translation [5]. Plagiarism detection is influenced by the type of plagiarism, it can be seen in previous research in this field. The previous researcher had been proposed technique to detect plagiarism including mono language or cross language, such as substring matching, keyword similarity, term

sequence matching with digital digests that known as fingerprint [6]. The fingerprint can be compute by digital digest through hashing function. This function will generate a collection numbers as a fingerprint, for example of hashing function is MD-5 hashing, it is one of most popular to be used, introduced by Stein and Meyer [7], the others are the Rolling Hash or SHA1 [8].

In Indonesian Language, there are some research that used the fingerprint method in plagiarism detection either with winnowing or just the origin of fingerprint method [9][10]. The winnowing algorithm used by Kusnawan et al in-plagiarism detection applications in Indonesian uses only simple pre-processing text such as removing spaces, and unneeded characters. This kind of method only can detect plagiarism in simple way but it cannot detect the more complex form such as changing active to passive voice. With addition stemming process in pre-processing that can help to detect this kind of plagiarism form. The proposed approach uses the winnowing algorithm in addition to the pre-processing stage that affects time processing. To overcome processing time, in pre-processing and processing processes requiring large quantities of document checking can utilize CPU capabilities that currently can consist of multiple cores to perform multiple processes simultaneously known as parallel processing. According to Khatter, many current CPUs consist of 8 - 12 cores that can be upgraded for more efficient process [11]. Through some trials can be seen time that can be saved by using parallel processing that involves several cores in one process. By checking the plagiarism with the winnowing algorithm through parallel processing can provide better results with more optimal performance.

2. Plagiarism Detection Theory

2.1. A Winnowing Algorithm

The methodology used in performing plagiarism detection using the fingerprint method with the Winnowing algorithm. The fingerprint method is a method that represents a document with a set of numbers obtained through a hash function [12]. This method was developed with winnowing algorithm [13]. How it works winnowing can be seen in the following stages:

a. A collection of text is read from a document.
   For example: "There are a lot of food"
b. A collection of texts is a cleaning process of unrelated characters such as spaces and dots.
   For example: after the cleaning is done the set of text becomes " Therearealotoffood ".
c. Furthermore, the word separation is based on N-Grams.
   For example: The N-Grams used are 5, then after the word separation becomes: There herea erear reare earea areal realo ealot aloto lotof otoff toffo offoo ffood
d. By using a certain hashing function, each word is determined by the number of hash fuction.
   For example : 656 333 1582 3172 4022 9914 1425 656 2355 3051 333 8395 1524 3181
e. From the sequence of numbers hereinafter referred to as the hash value will be formed into several windows. Each window consists of the same number of numbers. A window can consist of 4 hash keys or another one.
   For example : Each window only contains 4 hash values , then the results will be like these :
   [656 **333** 1582 3172] [333 1582 3172 4022] [1582 3172 **4022** 9914] [3172 4022 9914 **1425**] [4022 9914 1425 **656**] [9914 1425 656 2355] [656 2355 3051 **333**] [2355 3051 333 8395] [3051 333 8395 1524] [333 8395 1524 3181]
f. From each window, the smallest hash key is chosen to form the fingerprint of a document. If there is the same smallest hash key in a different window, then it is taken that occupies the rightmost position.
   For example: 333 4022 1425 656 333 333

After do all the step above, the fingerprint that represent the example document are 333 4022 1425 656 333 333. It can be varied depend on the size of the document and the hashing method to be used to generate fingerprint

2.2. *Jaccard Coeficient*

The detection of plagiarism can be seen from the degree of similarity between documents. Measuring the degree of similarity between documents using Jaccard Similarity Coeficient [14] :

$$Jaccard\ Similarity\ (A,B) = \frac{P\ (A \cap B\ )}{P\ (A \cup B)} \quad (1)$$

For examples A and B are two documents. Measures of similarity between documents A and B are derived from the same number of features between document A and document B then divided by the total of all the features present in document A and document B.

3. Methodology

The methodology that been used in this research briefly can be seen in figure 1. The detail explanation from each stage can be seen below:

1. Preparing the Dataset

   The dataset is collected from the online library of Bina Nusantara University that contain e-thesis file of students of Computer Science with Multimedia and Game topics. The data collected consists of 100 documents that have different file sizes (See Table 1) and using Indonesian Language. As is known thesis consists of several chapters, and the most likely most plagiarism occurs is in chapter 1 which is the introductory chapter and chapter 2 which consists of theoretical basis used in the study.

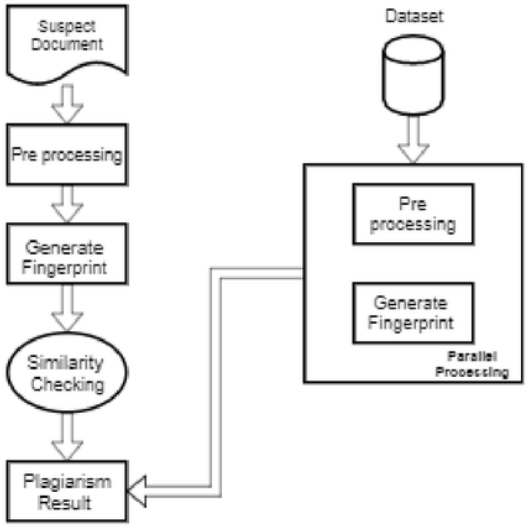

Figure 1. The stage of the Plagiarism Detection Technique

Table 1. Dataset Testing

No	Category	Quantity
1	Short Document (5 -10 pages)	49
2	Long Document (> 10 pages)	51
	Total Document	100

Treatment of the dataset to be tested consists of several treatments, namely the addition of the exact same sentence with another document, the change of sentence intonation from active sentence to passive sentence and there is also a document that does not receive any treatment so according to the original result of the document obtained (See Table 2).

Table 2. Some Treatment in Dataset

No	Treatment	Quantity
1	Original Document	40
2.	Add plagiarism 20% in Short Document	20
3.	Add plagiarism 40% in Long Document	20
4.	Active – Passive Changed	20
	Total Document	100

2. Generate Corpus
This activity generate corpus from the dataset that contain two processing: pre-processing and generate fingerprint for all documents in the dataset. In the pre-processing stage, the document must have docx extension and if the document has different extension then it converts first to docx extension. Then the all sentences in the document will be change in basic word through stemming process [15]. The next process is to generate fingerprint from the stemming result. Before generate fingerprint, all the white space or unrelated character will be removed from the stemming result. Separate the basic word using N-gram then generate number from the basic word using SHA1 hashing function that generate hash values Collect the hash value in each window then select the appropriate hash value. All the selected hash value knowns as the fingerprint from one document. The illustration from corpus generating can be seen in Figure 2. All the documents have been processed through parallel processing to saving processing time. This step to generate corpus only have to run one times as long the parameter in the plagiarism detection does not changed and can be used as reference corpus for any suspected document.

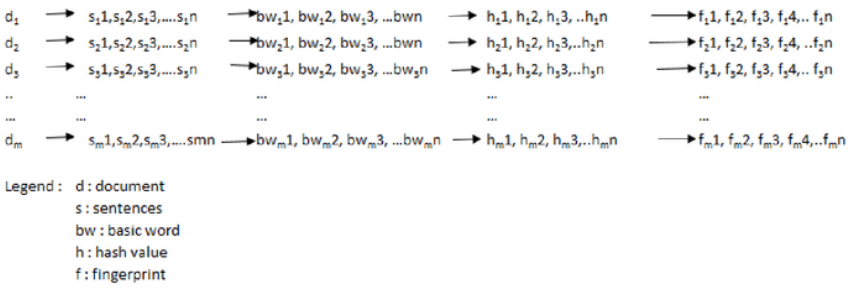

Figure 2. The illustration of each process in generating corpus

3. Checking of the Suspected Document
The suspected document that want to be checked must have docx extension and also do the preprocessing and generate fingerprint as same as generate corpus.
4. Similarity Checking
Identify plagiarism using the similarity checking with Jaccard Coeficient. The fingerprint from the suspected document checked with the fingerprint from the corpus through the similarity

checking and the result from this similarity checking represented in percentage to define how similar between both documents.

5. Plagiarism Result
   Based on the result from similarity checking then compared with the threshold 0.45 [9]. If the similarity degree is more than 45% , the suspected document has plagiarism, and if less than equal with 45% ,the suspected document does not have plagiarism.

4. Result and Discussion

4.1. Experiment Result

The experiment consists of two parts: test on various windows and 5 N-Gram, then test on process done without multiprocessing and with multiprocessing. Here are the test results by trying different window variations that are used to determine the level of similarity (Table 3):

Table 3. Similarity Degree related with the Window Size and 5 N-Grams

[Gram,Window Size]	[5,5]	[5,10]	[5,20]	[5,30]
Document 1	**47.66%**	**46.58%**	43.39%	42.36%
Document 2	**47.32%**	**45.02%**	41.95%	37.96%
Document 3	**46.60%**	44.37%	40.80%	37.86%
….	…	…	…	…
….	…..	….	….	…
Document 100	26.79%	25.87%	23.17%	19.89%

From the test results seen with the same N-Gram that is 5 N-Gram with various window variations show the greater the window size, the value of similarity is decreasing. Therefore, it can be concluded that the larger window size decreases the similarity level between the corpus document and the document suspected. In the winnowing algorithm if the window size is 10 means from all the fingerprint in one document will be grouped by ten in one window then the fingerprint with the smallest number be selected in each window. It can decrease the chance to be the selected fingerprint as compare with the size of window is 5. The more bigger size of the window has less selected fingerprints than the smallest size of the window. With more selected fingerprint can increase the chance to be compare with the fingerprint of suspected document.

In table 3 document 1 has the highest level of similarity compared to other documents when compared with suspected documents. According to threshold 0.45 then it can be concluded that the suspected document is a plagiarism document because the similarity degree with document 1,2 and document 3 more than 45% with window size is 5 and 5 N-Gram.

For the second stage of testing comparing processing time between parallel processing processes and without parallel processing (See Table 4). The process tested in the second stage is the stage of generating a fingerprint from Corpus that has 100 documents. The core process that is owned on the laptop used for testing consists of 4 suspected cores.

Table 4. Comparison of Processing Time

	Without Parallel Process	**With Parallel Process**	**Saving Time**
Time Processing	208.7894 Minutes	52.9052 Minutes	74,66%

Based on table 4, the process fingerprint that using parallel processing takes 52.9052 minutes and tit save processing time around 74.66%. If we cannot use the GPU's processor, this kind of technique is one of solution to process a lot of document more quickly.

5. Conclusion

The detection of plagiarism in Indonesian Texts can be done with Winnowing preceded by a pre-processing stemming stage. Although the addition of this stage affects time processing, but it can save time with parallel processing. The appropriate number of hash values for winnowing based on test results is to use 5 hash values on each winnow. With winnow the greater the effect on the level of similarity between documents.

For the further research, this plagiarism detection technique will be improved not only in Indonesian Language, but cross language between Indonesian with English. As we know there are more English text or documents that provide online in internet.

References

[1] Parker K, Lenhart A,and Moore K 2011 The Digital Revolution and Higher Education *Pew Res.* **202** 29.
[2] Maurer HA, Kappe F, Zaka B 2006 Plagiarism-a survey. *J. UCS.* **12(8)** 1050-84.
[3] Joy M and Luck M 1999 Plagiarism in programming assignments *IEEE Trans. Edu.* **42(2)** pp 129–33.
[4] Pupovac V and Fanelli D 2015 Scientists Admitting to Plagiarism: A Meta-analysis of Surveys *Sci. Eng. Ethics* **21(5)** pp 1331–52.
[5] Alzahrani SM, Salim N, Abraham A and Member S 2012 Understanding Plagiarism Linguistic Patterns, Textual Features, and Detection Methods *IEEE Trans. Syst. Man Cybernetics Part C (Applications and Reviews)* **42(2)** 133-49.
[6] Stein B and zu Eissen SM 2007 Fingerprint-based Similarity Search and its Applications *Forsch. und wissenschaftliches Rechn.* pp 85–98.
[7] Stein B and Meyer S 2006 Near Similarity Search and Plagiarism Analysis *Data Inf. Anal. Knowl. Eng.* 430-7.
[8] Wang J, Shen HT, Song J and Ji J 2014 Hashing for similarity search: A survey *arXiv preprint arXiv:1408.2927*.
[9] Wibowo AT, Sudarmadi KW and Barmawi AM 2013 Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents *IEEE Proc. ICoICT* pp 128-33.
[10] Kusmawan PY, Yuhana UL and Purwitasari D 2009 Aplikasi Pendeteksi Penjiplakan pada File Teks dengan Algoritma Winnowing pp 1–11.
[11] Khatter H and Aggarwal V 2014 Efficient parallel processing by improved CPU-GPU interaction *Proc. ICICT* pp 159–61.
[12] Meuschke N and Gipp B 2013 State-of-the-art in detecting academic plagiarism *Int. J. Educ. Integr.* **9(1)** pp 50–71.
[13] Schleimer S, Wilkerson DS, Aiken A and Berkeley UC 2003 Winnowing : Local Algorithms for Document Fingerprinting *Proc. ACM SIGMOD Int. Conf. Manage. Data* pp 76-85.
[14] Niwattanakul S, Singthongchai J, Naenudorn E and Wanapu S 2013 Using of Jaccard Coefficient for Keywords Similarity *Int. MultiConference Eng. Comput. Sci.* vol I pp 380–4.
[15] Librarian A and Kukuh R 2017 Sastrawi [Online]. Available: https://github.com/sastrawi/sastrawi.

Check Plagiarism Detection for Indonesian Language.pdf

ORIGINALITY REPORT

9%	8%	7%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Universitas Negeri Surabaya The State University of Surabaya Student Paper	6%
2	research.aalto.fi Internet Source	2%
3	china.iopscience.iop.org Internet Source	1%
4	R Bahana, F L Gaol, T Wiguna, S W H L Hendric, B Soewito, E Nugroho, B P Dirgantoro, E Abdurachman. "Performance test for prototype game for children with adhd", Journal of Physics: Conference Series, 2018 Publication	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		